Geophysical Research Abstracts
Vol. 15, EGU2013-13222-1, 2013
EGU General Assembly 2013

Application of large-scale computing infrastructure for diverse environmental research applications using GC3Pie

Sergio Maffioletti (1), Nicholas Dawes (2), Mathias Bavay (2), Sofiane Sarni (3), Michael Lehning (2,3)

(1) University of Zurich, Switzerland, (2) WSL Institute for Snow and Avalanche Research, SLF, Switzerland (dawes@slf.ch), (3) Ecole Polytechnique Federale de Lausanne, EPFL, Switzerland

The Swiss Experiment platform (SwissEx: http://www.swiss-experiment.ch) provides a distributed storage and processing infrastructure for environmental research experiments. The aim of the second phase project (the Open Support Platform for Environmental Research, OSPER, 2012-2015) is to develop the existing infrastructure to provide scientists with an improved workflow. This improved workflow will include pre-defined, documented and connected processing routines. A large-scale computing and data facility is required to provide reliable and scalable access to data for analysis, and it is desirable that such an infrastructure should be free of traditional data handling methods. Such an infrastructure has been developed using the cloud-based part of the Swiss national infrastructure SMSCG (http://www.smscg.ch) and Academic Cloud.

The infrastructure under construction supports two main usage models:
1) Ad-hoc data analysis scripts: These scripts are simple processing scripts, written by the environmental researchers themselves, which can be applied to large data sets via the high power infrastructure. Examples of this type of script are spatial statistical analysis scripts (R-based scripts), mostly computed on raw meteorological and/or soil moisture data. These provide processed output in the form of a grid, a plot, or a kml.
2) Complex models: A more intense data analysis pipeline centered (initially) around the physical process model, Alpine3D, and the MeteoIO plugin; depending on the data set, this may require a tightly coupled infrastructure. SMSCG already supports Alpine3D executions as both regular grid jobs and as virtual software appliances. A dedicated appliance with the Alpine3D specific libraries has been created and made available through the SMSCG infrastructure. The analysis pipelines are activated and supervised by simple control scripts that, depending on the data fetched from the meteorological stations, launch new instances of the Alpine3D appliance, execute location-based subroutines at each grid point and store the results back into the central repository for post-processing. An optional extension of this infrastructure will be to provide a 'ring buffer'-type database infrastructure, such that model results (e.g. test runs made to check parameter dependency or for development) can be visualised and downloaded after completion without submitting them to a permanent storage infrastructure.

Data organization
Data collected from sensors are archived and classified in distributed sites connected with an open-source software middleware, GSN. Publicly available data are available through common web services and via a cloud storage server (based on Swift). Collocation of the data and processing in the cloud would eventually eliminate data transfer requirements.

Execution control logic
Execution of the data analysis pipelines (for both the R-based analysis and the Alpine3D simulations) has been implemented using the GC3Pie framework developed by UZH. (https://code.google.com/p/gc3pie/). This allows large-scale, fault-tolerant execution of the pipelines to be described in terms of software appliances. GC3Pie also allows supervision of the execution of large campaigns of appliances as a single simulation. This poster will present the fundamental architectural components of the data analysis pipelines together with initial experimental results.